Global Elo Model for European Football*

A Multi-Season, Cross-League Rating and Prediction Framework

John Zhang

December 18, 2025

Table of contents

*Code and data are available at: https://github.com/Clearsky21z/Soccer_Elo_Model

1 Introduction

Measuring club strength across European football leagues is difficult. Domestic league tables reflect performance within a single competition and season, while UEFA country and club coefficients aggregate results across tournaments but adjust slowly and are not directly interpretable as team-level strengths. For tasks such as forecasting match outcomes or comparing clubs across leagues, a single, dynamically updated rating system is preferable—provided that it can place teams from the "Big Five" leagues (England, Spain, Italy, Germany, France) on a common scale, respond to new information in the current season, incorporate historical performance while discounting older seasons, and be tuned using a proper predictive scoring rule rather than rule-of-thumb choices.

This paper develops a global multi-season Elo model for domestic league matches in the Big Five leagues from the 2015/16 season up to the 2025/26 season (cut off at 15 December 2025). The specification builds on the Elo framework used in sports analytics and football modelling, as described in Beggs (2024) and Egidi, Karlis, and Ntzoufras (2026), and extends it in three main directions. First, season-specific UEFA country coefficients are used as league-level offsets so that teams from stronger leagues are, on average, rated higher than teams from weaker leagues. Second, ratings are shrunk toward a global baseline between seasons, and newly promoted teams start below that baseline to reflect the typical gap between incumbent top-flight clubs and new entrants. Third, older seasons are down-weighted through an explicit relevance function, and rating updates for the current season use an inflated K-factor so that ratings respond more strongly to recent results.

Model hyperparameters are chosen by minimising a relevance-weighted Brier score for match outcomes on a held-out test set, following the probability-forecast evaluation approach introduced by Brier (1950), and using a box-constrained quasi-Newton optimiser. The final output is an end-of-sample global ranking of clubs in Elo units, together with an estimate of out-of-sample predictive accuracy.

To conduct the analysis in a transparent and reproducible way, the workflow is implemented entirely in R and organised around a small set of well-established packages. Raw match files and intermediate tables are imported with `readr` (Wickham, Hester, and Bryan (2024)), and project-relative file paths are handled with `here` (Müller (2020)). Dates and team-name text fields are standardised using `lubridate` (Grolemund and Wickham (2011)) and `stringr` (Wickham (2023)). Data cleaning, reshaping, and aggregation rely on `dplyr` (Wickham et al. (2023)) and `tidyr` (Wickham, Vaughan, and Girlich (2024)), with compact summary tables supported by `tibble` (Müller and Wickham (2023)). Figures are produced with `ggplot2` (Wickham (2016)), with axis/scale formatting handled by `scales` (Wickham, Pedersen, and Seidel (2025)) and label placement supported by `ggrepel` (Slowikowski (2024)). Tables are generated dynamically using `knitr` (Xie (2014)), ensuring that descriptive summaries and model outputs update automatically when the underlying data or tuning results change.

The fitted model yields a calibrated but intentionally simple probabilistic baseline for Big-5 domestic outcomes. In the final refit, the relevance-weighted tuning selects a moderate update size and strong recency weighting, and the resulting overall Brier score is approximately 0.159 over the full sample. The final rating table contains 163 clubs, with end-of-sample Elo ratings spanning roughly 860 to 1402. At the top of the ranking, Bayern Munich is highest-rated (about 1402), followed by Paris SG, Real Madrid, Barcelona, and Arsenal, while Dortmund appears in the top 25 (rank 13), consistent with sustained upper-tier Bundesliga performance.

These results indicate that a domestic-only, multi-season Elo model—augmented with season-to-season shrinkage, a promotion penalty, and recency emphasis—can produce a coherent cross-league ordering on a single scale while maintaining interpretable match-level probabilities. At the same time, the analysis also motivates a careful comparison with widely used public rating systems (FootballDatabase and Opta), which is presented in Appendix {#sec-comparison} using aligned within-set ranks.

The remainder of this paper is structured as follows. Section 2 describes the data and preprocessing. Section 3 introduces the global multi-season Elo model, including shrinkage, promotion handling, and recency weighting. Section 4 reports predictive performance and presents the final club rankings. Section 5 discusses interpretation, limitations, and directions for future work and finally Section A describes detailed data retrieval process and comparison with established football rating systems

2 Data

The cleaned dataset spans the 2015/16 season through the 2025/26 season, with the sample cut off at 15 December 2025 (i.e., the 2025/26 season is partial). Each row corresponds to one match and includes the match date, league identifier, home and away team names, full-time home and away goals (FTHG/FTAG), and a full-time result indicator ($Y \in H, D, A$) for home win, draw, or away win.

To demonstrate how the cleaned match table can be filtered into interpretable team-specific summaries, Table 1 lists Borussia Dortmund's 10 most recent home matches in the dataset. Table 2 aggregates all Dortmund home matches in the cleaned file into a compact performance summary (wins/draws/losses and average goals for/against), providing an example of how match-level records translate into descriptive performance statistics.

Table 2: Summary statistics for all Dortmund home matches

Matches	Wins (H)	Draws (D)	Losses (A)	Avg goals for	Avg goals against	Avg total goals
175	123	28	24	2.66	1.14	3.79

Table 1: Most recent 10 home matches for Dortmund

Date	Season	League	Opponent	Score	Result	Total goals
2025-11-22	2025/26	Germany	Stuttgart	3–3	D	6
2025-10-25	2025/26	Germany	FC Koln	1–0	H	1
2025-10-04	2025/26	Germany	RB Leipzig	1–1	D	2
2025-09-21	2025/26	Germany	Wolfsburg	1–0	H	1
2025-08-31	2025/26	Germany	Union Berlin	3–0	H	3
2025-05-17	2024/25	Germany	Holstein Kiel	3–0	H	3
2025-05-03	2024/25	Germany	Wolfsburg	4–0	H	4
2025-04-20	2024/25	Germany	M'gladbach	3–2	H	5
2025-03-30	2024/25	Germany	Mainz	3–1	H	4
2025-03-08	2024/25	Germany	Augsburg	0–1	A	1

Table 3: Match counts and average goals per match by league

League	Matches	Avg total goals	Avg home goals	Avg away goals
England	3920	2.83	1.55	1.27
France	3677	2.71	1.51	1.20
Germany	3168	3.06	1.70	1.36
Italy	3930	2.74	1.49	1.25
Spain	3940	2.62	1.49	1.13

To provide transparent sample coverage, Table 3 reports the total number of matches available per league in the cleaned file, along with average home goals, away goals, and total goals per match.

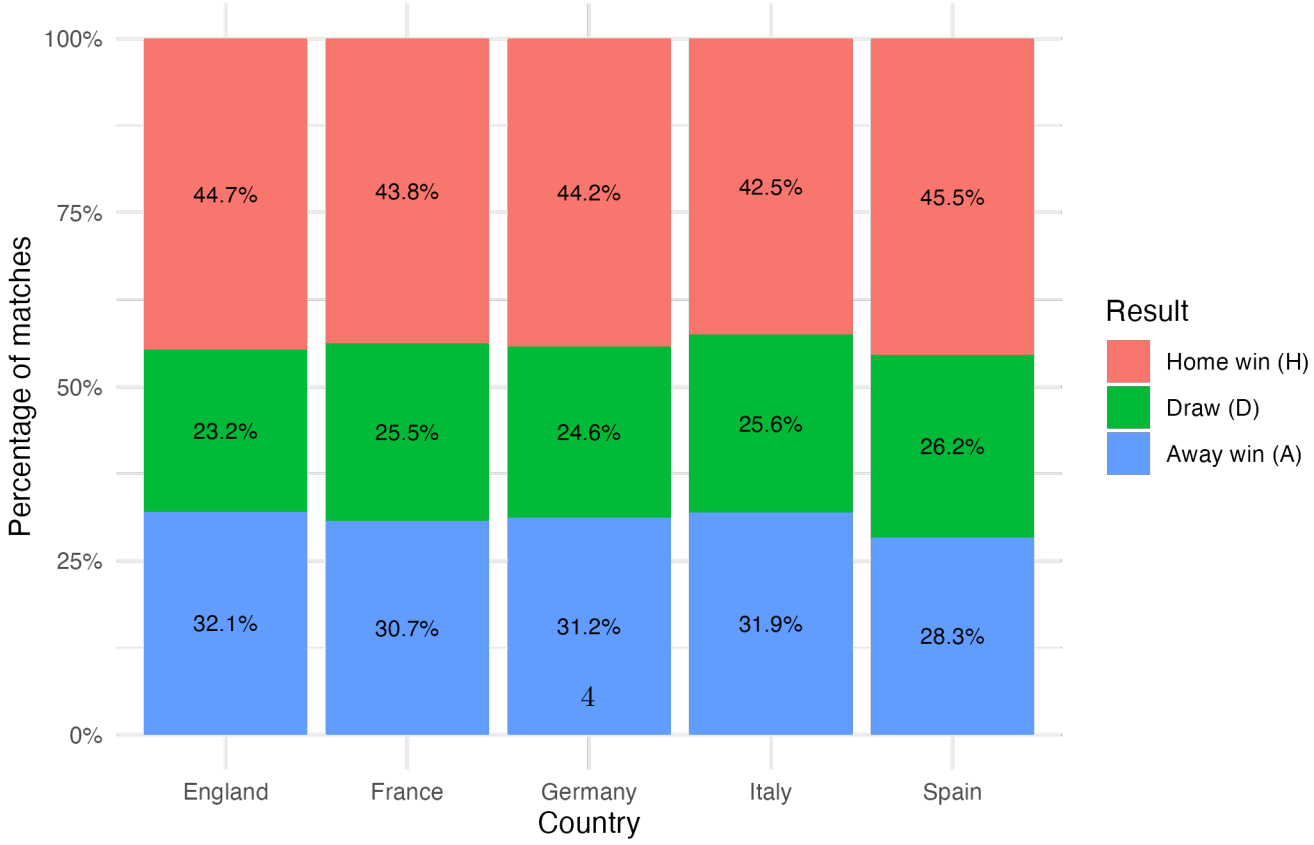

Figure 1: Home/draw/away result proportions by league

home and away teams in Figure 2, which provides a direct view of how frequently common scoreline components (e.g., 0–1–2 goals) occur.

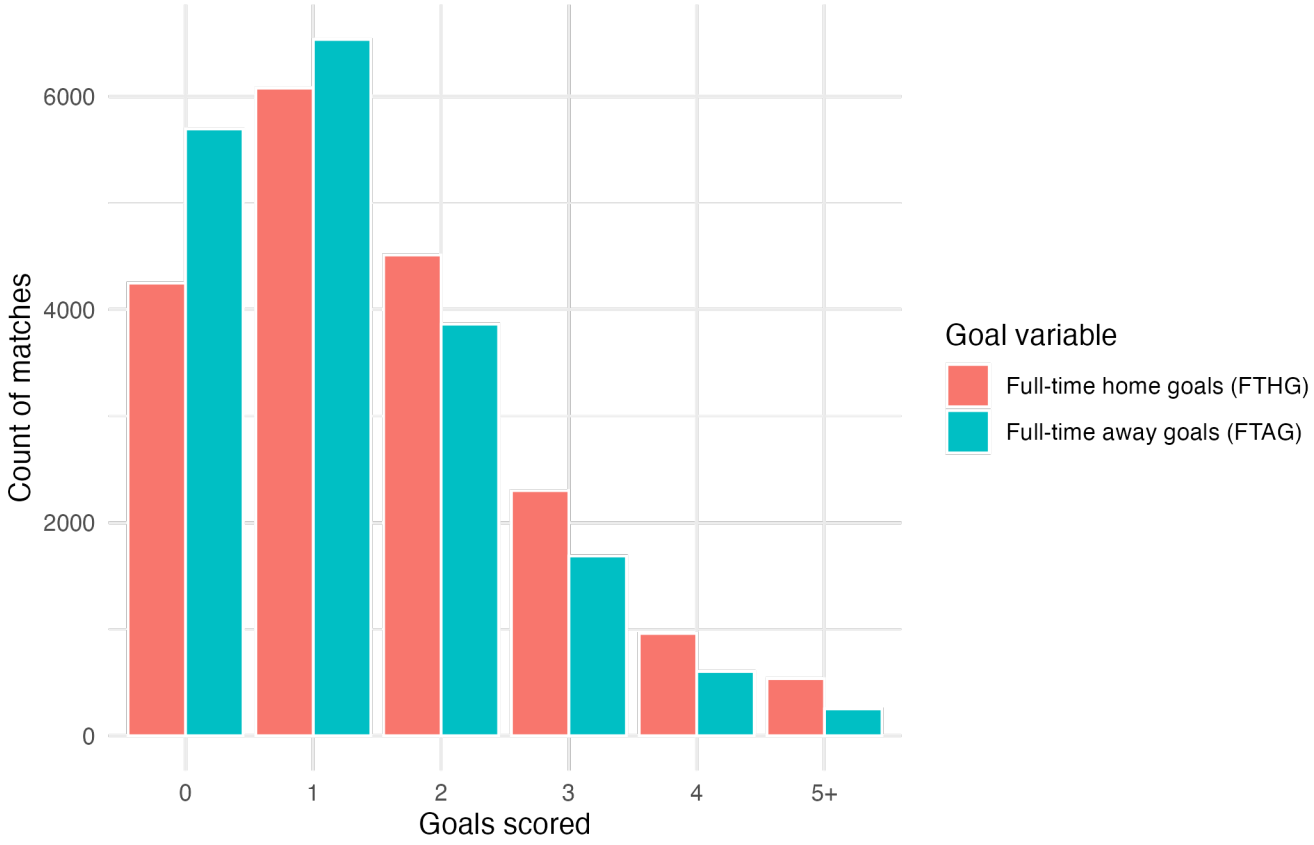

Figure 2: Distribution of home and away goals per match

To connect domestic league results to broader cross-league strength, matches are merged with a league–season table of UEFA country coefficients. Let $(c_{\ell,s})$ denote the coefficient for league (ℓ) in season (s). This coefficient is rescaled into a league strength offset $(\delta_{\ell,s})$, which enters the Elo expected-score calculation as an additive league–season term. The time variation in $(c_{\ell,s})$ across the Big-5 is shown in Figure 3.

After preprocessing, each match (i) is represented by the season (s_i), start year (t_i), league (ℓ_i), home and away teams (h_i) and (a_i), outcome $(Y_i \in H, D, A)$, and the UEFA-based offset (δ_{ℓ_i,s_i}). The ordered collection $((h_i, a_i, s_i, t_i, \ell_i, Y_i, \delta_{\ell_i,s_i})_{i=1}^{N})$ is the input to the global Elo model described in Section 3.

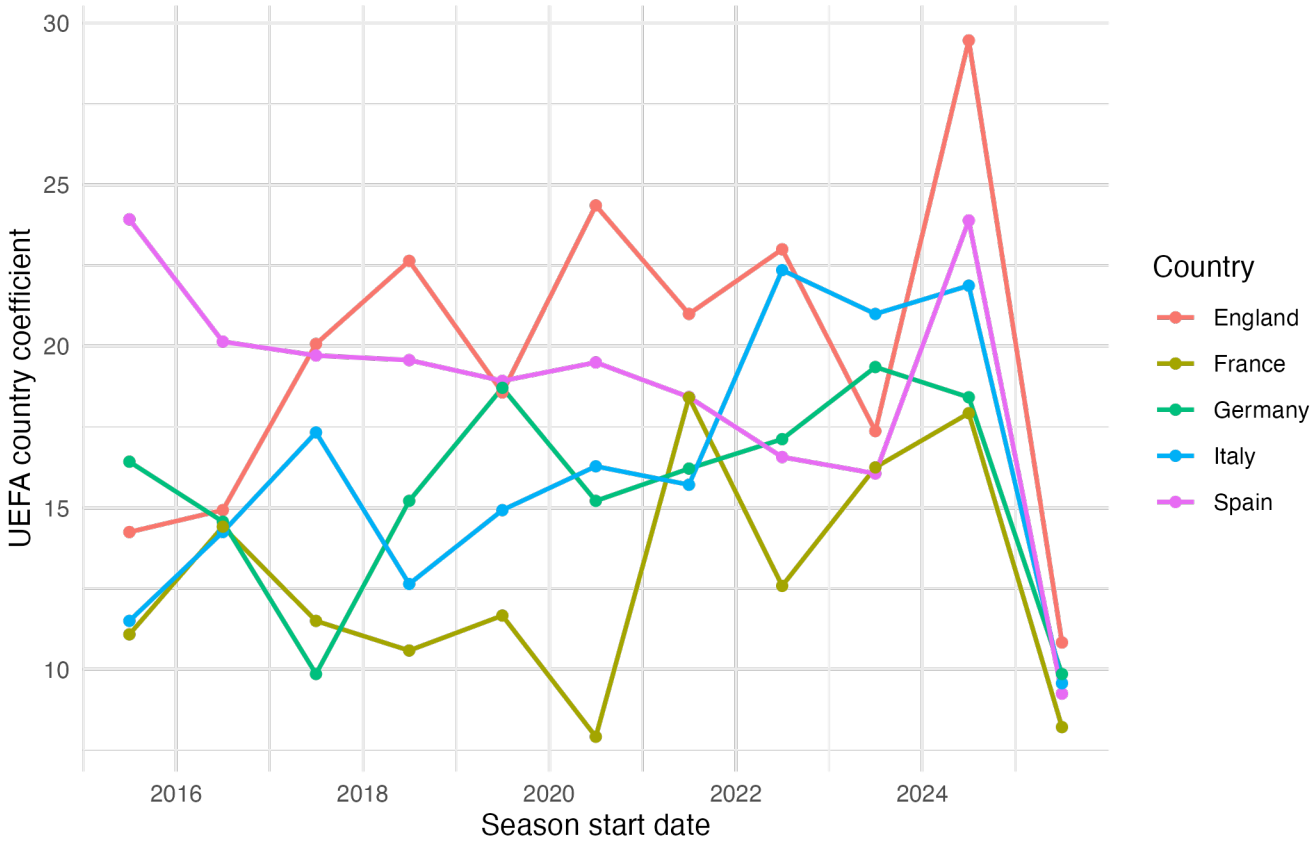

Figure 3: UEFA league coefficients over time

3 Elo rating model

The model assigns to each team j at match index i an Elo rating $R_{j,i}$ measured in points. At the beginning of the dataset all teams share a common baseline rating $R_0 = 1200$. For a given match i between home team h_i and away team a_i, with pre-match ratings $R_{h_i,i}$ and $R_{a_i,i}$, the model incorporates a home advantage H (fixed at 100 Elo points) and a league offset δ_{ℓ_i,s_i} common to both teams, derived from the UEFA coefficient.

The expected home "score" is defined using the standard Elo logistic link

$$E_{h_i,i} \;=\; \frac{1}{1+10^{-(R_{h_i,i}+H+\delta_{\ell_i,s_i}-(R_{a_i,i}+\delta_{\ell_i,s_i}))/400}}.$$

Because domestic matches involve two teams from the same league in the same season, the offset δ_{ℓ_i,s_i} cancels algebraically inside the parentheses; however, it remains in the formulation so the framework can be extended to cross-league fixtures where the offsets differ.

The observed full-time result Y_i is encoded as a home score

$$S_{h_i,i} = \begin{cases} 1 & \text{if } Y_i = \text{H (home win)}, \\ 0.5 & \text{if } Y_i = \text{D (draw)}, \\ 0 & \text{if } Y_i = \text{A (away win)}. \end{cases}$$

The away score is $S_{a_i,i} = 1 - S_{h_i,i}$. After match i, ratings are updated according to

$$R_{h_i,i+1} \;=\; R_{h_i,i} + K_{\text{eff},i}\,(S_{h_i,i} - E_{h_i,i}),$$

$$R_{a_i,i+1} \;=\; R_{a_i,i} + K_{\text{eff},i}\,(S_{a_i,i} - (1 - E_{h_i,i})),$$

where $K_{\text{eff},i}$ is a match-specific effective K–factor that controls how quickly ratings react to the result. Ratings for other teams remain unchanged at time $i+1$.

3.1 Multi-season dynamics: shrinkage and promoted teams

To prevent ratings from drifting over long time horizons and to reflect changes in team quality between seasons, two mechanisms are applied at season boundaries. Let s index seasons and $t(s)$ be the start year of season s (for example, $t(2015/16) = 2015$). At the end of season s, suppose team j has rating $R_{j,\text{end}(s)}$. Before the first match of season $s+1$, its carried-over rating is shrunk toward the global baseline R_0 via

$$R^{\text{carried}}_{j,\text{start}(s+1)} \;=\; R_0 + \lambda\,(R_{j,\text{end}(s)} - R_0),$$

where $0 \le \lambda \le 1$ is a shrinkage parameter. The case $\lambda = 1$ corresponds to full carry-over with no regression to the mean, $\lambda = 0$ resets all teams to the baseline at the start of each season,

and intermediate values partially preserve past performance while pulling ratings back toward R_0. Teams present in the previous season use $R^{\text{carried}}_{j,\text{start}(s+1)}$ as their starting rating.

For newly promoted teams or teams that appear for the first time in the dataset in season s, there is no carried-over rating. These teams start below the baseline at

$$R_{j,\text{start}(s)} = R_0 - P,$$

where $P \geq 0$ is a promotion penalty. This reflects the typical gap between incumbent top-division clubs and teams promoted from lower divisions. Thus, at the start of each season, every team either inherits a shrunken rating from the previous season or is treated as new or promoted and assigned the penalised starting rating $R_0 - P$.

3.2 Recency and current-season emphasis

Historical results are informative, but matches from many seasons ago should have less influence than recent results. This is handled through relevance weights in the loss function and a current-season adjustment to the K–factor.

Let T denote the start year of the most recent season in the dataset, and let t_i be the start year of the season containing match i. Define the season age of match i as

$$\text{age}_i = T - t_i.$$

A relevance weight is then defined as

$$w_i = \rho^{\,\text{age}_i},$$

where $0 < \rho < 1$ is a relevance decay parameter. Smaller values of ρ down-weight older seasons more aggressively. These weights do not affect the Elo update itself; instead, they are used when evaluating model fit and tuning the hyperparameters (Section **??**). Matches from the most recent season have $\text{age}_i = 0$ and thus $w_i = 1$, while matches from older seasons have $w_i < 1$.

To emphasise the 2025/26 season in the rating dynamics, the effective K–factor depends on whether a match belongs to the most recent season. Specifically,

$$K_{\text{eff},i} = \begin{cases} K \cdot c, & \text{if } t_i = T, \\ K, & \text{if } t_i < T, \end{cases}$$

where $K > 0$ is the base K–factor and $c \geq 1$ is a current-season multiplier. For matches in the most recent season, ratings therefore move c times as much in response to the term $(S_{h_i,i} - E_{h_i,i})$ as they do in earlier seasons. Together, the relevance weights $\{w_i\}$ and the current-season multiplier c ensure that older matches are discounted when assessing predictive performance and that ratings are more sensitive to current-season outcomes.

3.3 Hyperparameter estimation via relevance-weighted Brier score

The global Elo model involves five main hyperparameters: the base K–factor K controlling the magnitude of rating updates; the inter-season shrinkage parameter λ; the promotion penalty P applied to new teams; the relevance decay parameter ρ for down-weighting old seasons; and the current-season K multiplier c. These hyperparameters are estimated by minimising a relevance-weighted Brier score on a held-out test set.

The N matches are first partitioned randomly into a training set $\mathcal{T}_{\text{train}}$ containing 70% of matches and a test set $\mathcal{T}_{\text{test}}$ containing the remaining 30%, using a fixed random seed. For any proposed parameter vector $\theta = (K, \lambda, P, \rho, c)$, the Elo update is applied sequentially over all matches, yielding for each match i an expected home score $E_{h_i,i}$ and an observed home score $S_{h_i,i}$. Relevance weights $w_i = \rho^{\text{age}_i}$ are then computed based on the season age. The relevance-weighted Brier scores on the training and test sets are given by

$$\text{Brier}_{\text{train}}(\theta) = \frac{\sum_{i \in \mathcal{T}_{\text{train}}} w_i (S_{h_i,i} - E_{h_i,i})^2}{\sum_{i \in \mathcal{T}_{\text{train}}} w_i},$$

and

$$\text{Brier}_{\text{test}}(\theta) = \frac{\sum_{i \in \mathcal{T}_{\text{test}}} w_i (S_{h_i,i} - E_{h_i,i})^2}{\sum_{i \in \mathcal{T}_{\text{test}}} w_i}.$$

The objective is to find $\hat{\theta}$ that minimises $\text{Brier}_{\text{test}}(\theta)$ subject to box constraints that restrict parameters to plausible ranges (for example, $5 \leq K \leq 80$, $0.5 \leq \lambda \leq 1$, $0 \leq P \leq 200$, $0.1 \leq \rho \leq 0.8$, $1 \leq c \leq 3$). The function $\text{Brier}_{\text{test}}(\theta)$ is treated as a black-box function of θ, and a box-constrained quasi-Newton optimisation algorithm (L-BFGS-B) is applied to obtain an approximate minimiser

$$\hat{\theta} = (\hat{K}, \hat{\lambda}, \hat{P}, \hat{\rho}, \hat{c}).$$

3.4 Final model and club rankings

Using the estimated hyperparameters $\hat{\theta}$, the model is refit on the full dataset. This involves recomputing the season ages and relevance weights using $\hat{\rho}$, applying the Elo updates sequentially over all matches with $K = \hat{K}$, $\lambda = \hat{\lambda}$, $P = \hat{P}$ and $c = \hat{c}$, and obtaining for each match the expected home score $\widehat{E}_{h_i,i}$ and realised home score $S_{h_i,i}$. The overall Brier score of the final model is

$$\text{Brier}_{\text{overall}} = \frac{1}{N} \sum_{i=1}^{N} (S_{h_i,i} - \widehat{E}_{h_i,i})^2,$$

which summarises the average squared error of the predicted home score across all matches.

To produce a single rating for each club, the match-by-match post-update ratings are tracked over time and the *final* Elo for club j is defined as its **last observed post-match rating** within the sample window. Let $i_j^\star = \max, i : j \in h_i, a_i$, denote the index of the most recent match in the dataset in which club j appears (either as home or away). The club's final Elo is then $\widehat{R} * j = R * j, , i_j^{\star\text{after}}$, i.e., the post-match rating immediately after its last recorded match. Clubs are ranked by sorting $\widehat{R}_j$ in descending order. This yields a global ranking across the Big Five leagues at the end of the sample period that reflects multi-season performance (with inter-season shrinkage and a promotion penalty) and stronger sensitivity to current-season outcomes through the current-season K multiplier.

4 Result

4.1 Hyperparameter estimates and predictive performance

Model hyperparameters were selected by minimising the relevance-weighted Brier score on a held-out 30% test set. The optimiser selected $\hat{K} \approx 21.8$, inter-season shrinkage $\hat{\lambda} \approx 0.888$, a promotion penalty $\hat{P} \approx 200$, relevance decay $\hat{\rho} \approx 0.10$, and a current-season multiplier $\hat{c} \approx 1.37$. These values imply (i) moderate match-to-match rating updates, (ii) meaningful regression toward the baseline between seasons, (iii) a substantial handicap for newly appearing/promoted clubs, and (iv) strong down-weighting of older seasons when tuning predictive performance.

Under the refit on the full dataset, the model's overall Brier score is approximately 0.159, summarising the mean squared error between the predicted home score $\widehat{E} * h_i, i$ and the encoded match outcome $S * h_i, i$ across all matches.

4.2 Distribution of global Elo ratings

The final Elo rating table contains 163 clubs. Summary statistics of the final Elo ratings are reported in Table 6. Ratings span from roughly 860 to 1402, with mean $\approx$ 1081 and standard deviation $\approx$ 127. This spread indicates substantial separation between the strongest and weakest clubs in the sample: differences of a few hundred Elo points correspond to large differences in implied expected score under the Elo logistic link.

4.3 Top-25 clubs and league composition

Table 4 reports the top 25 clubs by final Elo rating at the end of the sample window. Bayern Munich ranks first (Elo $\approx$ 1402), followed by Paris SG, Real Madrid, Barcelona, and Arsenal. Dortmund also appears in the top 25 (rank 13), consistent with sustained upper-tier Bundesliga performance over the sample period (Table 4).

To summarise which leagues contribute most to the top of the ranking, Table 5 counts the number of top-25 clubs by league. England and Italy each contribute six clubs, Spain contributes five, and France and Germany contribute four each (Table 5).

Table 4: Top 25 clubs by final Elo rating

Club	Country	Elo rating
Bayern Munich	Germany	1402
Paris SG	France	1374
Real Madrid	Spain	1372
Barcelona	Spain	1371
Arsenal	England	1359
Inter	Italy	1335
Ath Madrid	Spain	1321
Napoli	Italy	1320
Villarreal	Spain	1312
Milan	Italy	1306
Man City	England	1303
Leverkusen	Germany	1299
Dortmund	Germany	1286
Roma	Italy	1280
Marseille	France	1271
Juventus	Italy	1270
Lens	France	1268
Liverpool	England	1264
RB Leipzig	Germany	1259
Aston Villa	England	1258
Chelsea	England	1258
Lille	France	1252
Bologna	Italy	1240
Betis	Spain	1227
Crystal Palace	England	1225

Table 5: Top-25 Elo clubs: Number of clubs by league

Country	Clubs in top 25
England	6
Italy	6
Spain	5
France	4
Germany	4

Table 6: Summary statistics of final Elo ratings

N clubs	Highest Elo	Lowest Elo	Mean Elo	Median Elo	SD Elo	IQR Elo
163	1402	860	1081	1067	127	199

5 Discussion

This paper constructs a single Elo-based rating system for clubs in the top divisions of England, Spain, Italy, Germany, and France using domestic league matches from 2015/16–2025/26 (up to the sample end date). Ratings update match-by-match with a fixed home advantage, carry across seasons via shrinkage toward a baseline ($R_0 = 1200$), and assign a promotion penalty to newly appearing clubs. Hyperparameters are chosen by minimising a relevance-weighted Brier score, placing extra emphasis on recent seasons.

With that model in place, the results highlight several broad patterns about Big-5 domestic football. One clear lesson is that team strength is strongly tiered: a small elite group sits above a dense cluster of strong contenders. In the end-of-sample ranking, Bayern Munich leads, followed closely by Paris SG, Real Madrid, Barcelona, and Arsenal (Table 4). This ordering implies that even among "top" clubs, there is meaningful separation, but also that many matches among the upper tier should be relatively competitive because the Elo gaps are not enormous.

A second takeaway concerns how top strength is distributed across leagues. The top-25 list is not evenly split by country: England and Italy contribute the largest shares, followed by Spain, with France and Germany contributing slightly fewer (Table 5). Interpreted literally, this suggests differences in the depth of elite clubs across the five leagues during the sample window.

To help interpret how these rankings align with established public systems, the paper also conducts a detailed comparison against FootballDatabase and Opta in Section A. That appendix shows where rank orders agree and where they diverge once the club set is aligned and rankings are interpreted within the same Big-5 subset.

At the same time, several limitations should shape how strongly we interpret these results. The model uses only categorical outcomes (win/draw/loss) and ignores match context such as injuries, squad rotation, travel, or managerial change, and it does not use goal margin (or xG) information. League strength enters via UEFA-based offsets rather than being estimated directly from cross-league match data. Predictive tuning relies on a random train/test split, which is convenient but not the same as forecasting forward through time. Finally, because the reported "final" club rating is the last observed post-match Elo at the cutoff date (not a peak-over-window measure), the ranking reflects end-of-sample strength more than a club's best point during the 2015/16–2025/26 period.

These constraints point naturally to future work. A time-ordered evaluation would better reflect real forecasting, and incorporating goal difference or expected-goals signals could reduce noise and improve calibration. Extending the dataset to include European competition matches would also allow league effects to be learned more directly rather than imposed through external coefficients. Lastly, reporting rolling or end-of-season ratings (and ideally uncertainty bands) would help distinguish persistent club quality from short-term form swings.

A Appendix

A.1 Data Retrieval Process

A.2 Comparison with Established Football Rating Systems

This appendix situates the global Elo model developed in this paper within the broader landscape of football rating methodologies. The comparison focuses on three widely referenced systems: (i) the *FootballDatabase World Club Elo Ratings*, (ii) the *FIFA Men's World Ranking*, and (iii) the *Opta Power Rankings*. Although these systems are all related—directly or indirectly—to Elo-style ideas (expected result plus rating updates after matches), they differ in scope, match weighting, use of score margin, season treatment, and the extent to which tuning is driven by predictive calibration. These differences help explain why the same club can occupy different positions across published rankings.

A.2.1 Conceptual scope and rated entities

This paper (Big-5 model). The model is restricted to domestic top-division league matches in England, Spain, Italy, Germany, and France (2015/16–2025/26). The goal is a *comparable* measure of club strength within a controlled setting: same match type (league play), same region (Big Five leagues), and consistent season structure.

FootballDatabase. FootballDatabase publishes a worldwide club ranking that aggregates domestic top tiers across many countries and also incorporates international club competitions. Its design goal is a single "world table" for clubs across leagues and continents, constructed using an Elo-style update with explicit competition weights and goal-difference scaling.

FIFA. FIFA ranks national teams rather than clubs, based on official international fixtures. The post-2018 method is commonly presented in Elo-like form with a match-importance factor; the rated entities and match calendar are fundamentally different from club football.

Opta. Opta's Power Rankings are also a worldwide club ranking. Opta describes the system as Elo-based, but implemented through a hierarchy of ratings (team, league, country, continent) so that results can propagate between levels when clubs play cross-league or cross-continent matches. Opta also reports transforming the internal rating to a 0–100 scale for presentation.

A.2.2 Model structure and updating principles

This section describes the three club systems and the FIFA system using matched terminology: (i) expected result, (ii) update equation, (iii) match weighting, (iv) margin of victory, and (v) time/season treatment.

A.2.2.1 This model

Expected result. For match i with home team h_i and away team a_i, the expected home score is computed via the Elo logistic transform using the pre-match rating difference and a fixed home-advantage term H.

Update equation. Ratings update as

$$R_{h,i+1} = R_{h,i} + K_{\text{eff},i}, (S_{h,i} - E_{h,i}), \qquad R_{a,i+1} = R_{a,i} + K_{\text{eff},i}, (S_{a,i} - (1 - E_{h,i})),$$

with $S_{h,i} \in 1, 0.5, 0$ encoding win/draw/loss.

Match weighting. All domestic league matches are treated as the same match type; the only systematic reweighting in the *update* is the current-season multiplier applied through $K_{\text{eff},i}$.

Margin of victory. No goal-difference multiplier is used; the outcome enters only through the categorical result (win/draw/loss).

Season/time treatment. Ratings shrink toward a baseline between seasons (parameter λ), promoted/new teams start below baseline (penalty P), and historical seasons are down-weighted for *tuning* via relevance weights in the loss.

Calibration/tuning. Hyperparameters are selected by minimizing a relevance-weighted Brier score on a held-out set, so the primary objective is explicitly predictive (probability calibration), not only rank plausibility.

A.2.2.2 FootballDatabase

FootballDatabase describes a modified Elo update of the form

$$R_{\text{new}} = R_{\text{old}} + K, G, (W - W_e),$$

where $W \in 1, 0.5, 0$ is the match result and W_e is the expected result from the Elo logistic function.

Match weighting (competition importance). The factor K varies by competition and stage (e.g., higher in major international competitions; different constants across domestic leagues).

Margin of victory. FootballDatabase includes an explicit goal-difference multiplier G (e.g., $G = 1$ for a draw/one-goal win; larger values for bigger winning margins).

Season/time treatment and promoted teams. FootballDatabase describes initial ratings for newly promoted teams that depend on league strength or tier, rather than estimating promotion effects from a single unified scoring-rule framework within a fixed-scope dataset.

In short, FootballDatabase emphasizes global coverage and match-importance weighting (via K) plus margin-of-victory responsiveness (via G), whereas the Big-5 model in this paper emphasizes controlled scope and predictive calibration.

A.2.2.3 FIFA Men's World Ranking

Public descriptions of FIFA's post-2018 ranking present an Elo-like update:

$$P_{\text{new}} = P_{\text{old}} + I, (W - W_e),$$

where I is a match-importance factor (higher for major tournaments than friendlies). Unlike the club models above, the rated entities are national teams and the match set is international fixtures, so direct numerical comparison to club ratings is not meaningful even when the algebra is similar.

A.2.2.4 Opta Power Rankings

Opta describes its Power Rankings as Elo-based, but implemented through a hierarchy: team ratings sit within league ratings, within country ratings, within continent ratings. Match results exchange rating points at the appropriate level(s): domestic league matches primarily affect teams (and their league), while cross-league competitions allow rating to move between leagues/countries/continents through the hierarchy.

Margin of victory. Opta states that the margin of victory affects the size of the rating exchange (larger wins imply larger exchanges).

Scale. Opta reports transforming its internal rating into a 0–100 scale using a power transform and min–max scaling for interpretability and presentation.

Because Opta's published description is high-level, the exact parameter values (e.g., effective learning rates by competition) are not fully transparent, but the key architectural distinction is the hierarchical propagation mechanism and the standardized public-facing scale.

A.2.3 Practical comparability of Ranking Systems

Because FootballDatabase and Opta are global systems (mixing domestic and continental competitions across many countries), while this paper is intentionally Big-5 domestic-only, the absolute rating scales are not directly comparable. The defensible comparison is therefore at the level of rank order, after aligning the club set and using a consistent snapshot of rankings. In this appendix, the aligned club set is defined as the top 25 clubs in the Big-5 Elo ranking (Table 7). FootballDatabase and Opta ranks are taken as within-set ranks over the same Big-5 subset (i.e., ranks re-indexed within the aligned Big-5 universe rather than interpreted as worldwide ranks), so that "rank 1" always means "best within the Big-5 subset," not best globally.

A direct club-by-club comparison of the three rank lists is shown in Table 7. To make disagreements visually transparent, Figure 4 connects each club's position across systems; line crossings indicate rank-order reversals between methods.

Overall agreement is summarised in Table 8. Within the aligned top-25 club set, the Big-5 Elo ordering has Spearman rank correlation of 0.705 with FootballDatabase and 0.550 with Opta, indicating moderate agreement in broad ordering and tiering, with closer alignment to FootballDatabase than Opta. The corresponding Kendall tau values are 0.520 (FootballDatabase) and 0.433 (Opta), consistent with moderate concordance in pairwise ordering. In more interpretable units, the median absolute rank gap is 5 places versus FootballDatabase and 9 places versus Opta.

The largest rank gaps are listed in Table 9 and Table 10. These discrepancies are informative rather than "errors": FootballDatabase incorporates global competition weighting and margin-of-victory adjustments, so clubs whose international/competition-mix signal differs from domestic Big-5 league performance can shift substantially relative to a domestic-only Elo. Opta's ranking is designed to reflect a broader performance signal than win/draw/loss alone in its proprietary framework, so clubs can be repositioned when underlying performance quality and opponent adjustments diverge from realised domestic results. Overall, the disagreements shown in Figure 4 and Table 9–Table 10 are consistent with the fact that this paper measures domestic Big-5 league strength, while external systems incorporate broader global strength signals, even when comparisons are restricted to the same Big-5 club subset.

Table 7

Club	Country	My Model Rank	FDB Rank	Opta Rank
Bayern Munich	Germany	1	1	2
Paris SG	France	2	3	4
Real Madrid	Spain	3	8	10
Barcelona	Spain	4	4	5
Arsenal	England	5	2	1
Inter	Italy	6	9	9
Ath Madrid	Spain	7	6	13
Napoli	Italy	8	16	22
Villarreal	Spain	9	20	31
Milan	Italy	10	13	18
Man City	England	11	5	3
Leverkusen	Germany	12	14	16
Dortmund	Germany	13	11	14
Roma	Italy	14	15	23
Marseille	France	15	24	27
Juventus	Italy	16	19	25
Lens	France	17	34	36
Liverpool	England	18	12	7
RB Leipzig	Germany	19	26	30
Aston Villa	England	20	7	6
Chelsea	England	21	10	8

Table 7

Club	Country	My Model Rank	FDB Rank	Opta Rank
Lille	France	22	32	35
Bologna	Italy	23	30	32
Betis	Spain	24	25	38
Crystal Palace	England	25	17	12

Table 8: Rank-order agreement between the Big-5 Elo model and other rankings

Comparison	N (clubs with rank)	Spearman rho	Kendall tau	Median Absolute Rank Gap
My Model vs FootballDatabase	25	0.705	0.520	5
My Model vs Opta	25	0.550	0.433	9

Table 9: Largest within-set rank disagreements: Big-5 Elo vs FootballDatabase

Club	My Model Rank	FDB Rank	Absolute Rank Gap
Lens	17	34	17
Aston Villa	20	7	13
Villarreal	9	20	11
Chelsea	21	10	11
Lille	22	32	10

Table 10: Largest within-set rank disagreements: Big-5 Elo vs Opta

Club	My Model Rank	Opta Rank	Absolute Rank Gap
Villarreal	9	31	22
Lens	17	36	19
Napoli	8	22	14
Aston Villa	20	6	14
Betis	24	38	14

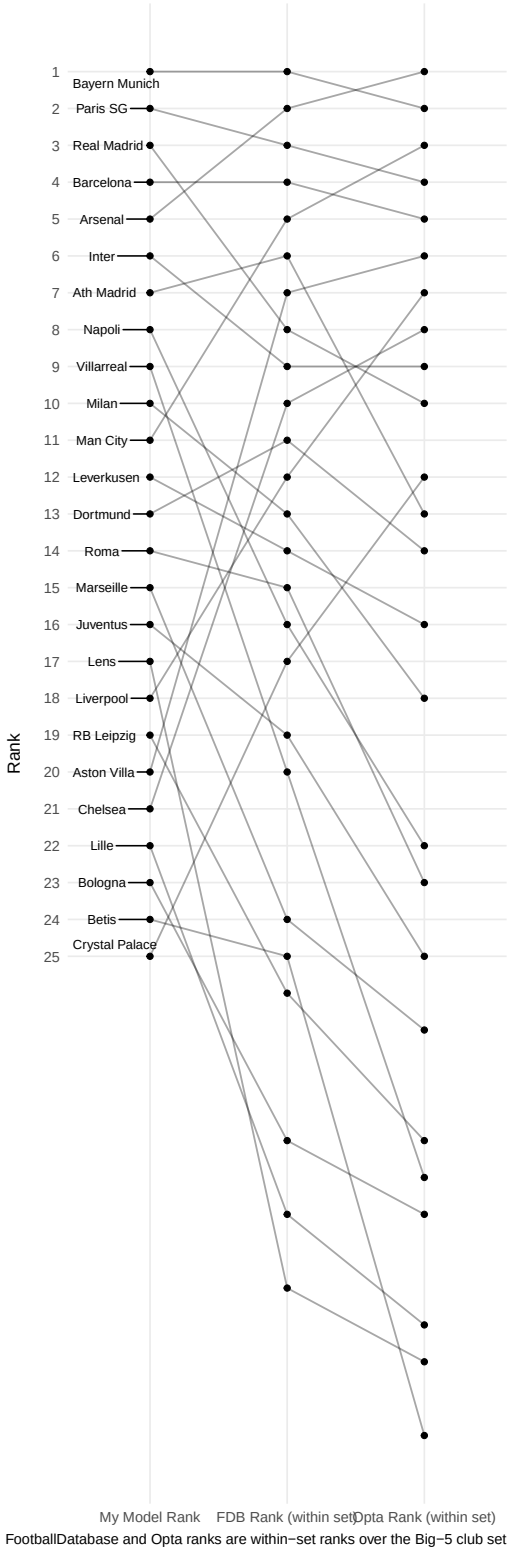

Figure 4: Rank comparison across systems for the aligned Big-5 league club set

References

Beggs, Clive. 2024. *Soccer Analytics: An Introduction Using r*. 1st ed. Chapman & Hall/CRC Data Science Series. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781003328568.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Egidi, Leonardo, Dimitris Karlis, and Ioannis Ntzoufras. 2026. *Predictive Modelling for Football Analytics*. 1st ed. Chapman & Hall/CRC Data Science Series. CRC Press. https://doi.org/10.1201/9781003186496.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Slowikowski, Kamil. 2024. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://CRAN.R-project.org/package=ggrepel.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2025. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.